Effects of energy- and climate policy in Germany

Többen, Johannes

Chapter 3

On the simultaneous Estimation of Physical and Monetary Commodity Flows•

3.1 Introduction

This paper considers the simultaneous estimation of commodity flows measured in both monetary and physical units from incomplete and partial information. Such problems arise regularly in the context of linking multisectoral and/or multiregional economic data, e.g., in Multiregional Input-Output (MRIO) tables, with data such as freight transportation accounts, material flow data, emission statistics or energy accounts in order to create integrated environmental-economic databases. Such databases have been used for a wide range of applications.

The consumption-based accounting of emissions, use of natural resources and the like (see inter alia Wiedmann, 2009; Lenzen et al., 2012a; Dietzenbacher et al., 2013; Lenzen et al., 2013; Tukker et al., 2013; Wood et al., 2014; 2015; Wiedmann et al., 2015) requires the estimation and harmonization of satellite accounts measured in physical units with backbone MRIOs measured in currency units. Recent efforts aim at constructing fully linked monetary and physical I–O databases. Examples include global databases, for example EXIOBASE2 (Wood et al., 2015), as well as regional physical-monetary I–O tables, for example for the city of Beijing (Zhang et al., 2014). In hybrid LCA applications, information about process requirements measured in physical units is frequently combined with monetary I–O accounts (see for example Joshi, 1999; Wiedmann et al., 2011; Arversen et al., 2013). MRIOs coupled with transportation models are employed for the prediction of future demand for passenger or freight transport and infrastructure planning (Zhao and Kockelmann, 2004; Ham et al., 2005; Kockelman et al., 2005; Caggiani et al., 2014).

Generally, one and the same commodity flow can be viewed from different perspectives, for example as the amount of tons transported from one region or sector to another or as the monetary value of the corresponding trade relationship. Thus, even if one is only interested in estimating unobserved commodity flows for a single dimension, using data measured in different units may provide important additional information in order to support the estimation process and to improve the quality of the

• This chapter is based on "On the simultanious estimation of physical and monetary commodity flows." published in Economic Systems Research (2017, 29:1). The paper was presented at the 24th International Input-Output Conference held in Seoul, Korea. The author would like to thank Jan Oosterhaven, Erik Dietzenbacher and the participants of the IIOA as well as three anonymous referees for their valuable comments and suggestions.

results. For example, freight transportation data may serve as an indirect source of information for the estimation of unobserved interregional trade flows (see, for example Kim et al., 1983; Jackson et al., 2006; Gallego and Lenzen, 2009; Park et al., 2009; Llano et al., 2010; Thissen et al., 2013). The other way round, monetary I–O tables may serve as an indirect source of information for the construction of physical I–O tables (Zhang et al., 2014; Wood et al., 2015).

In any case, the integration of data measured in different units and originating from a variety of sources requires overcoming several challenges:

- Combining data measured in monetary and physical units requires the access to or the estimation of price relationships.
- Monetary and physical commodity flow data are often published in different and possibly mismatching commodity classifications as well as at different levels of aggregation.
- Information is typically only available to a limited extent and/or is incomplete (e.g., due to suppressed data of low confidence or due to confidentiality), which leads to an underdetermined estimation problem, i.e., where the number of unknowns exceeds the number of data points.
- Resulting commodity flow accounts must adhere to joint financial, mass and/or energy balances.

In the applications cited above these challenges are addressed in a step-wise manner. Typically, these steps involve the imputation of missing data, the transformation into other units of measurement and the harmonization between different classifications and levels of aggregation.

This paper proposes a novel model based on the principles of *maximum entropy* that allows for estimating monetary and physical commodity flows simultaneously and consistent with financial and mass balances under the restrictions of partial and incomplete information, different levels of aggregation and mismatching commodity classifications. The *maximum entropy* principle to statistical inference was originally proposed by Jaynes (1957) as the least biased estimator on the basis of partial knowledge about the system under study.

Maximum entropy models and the closely related *minimal cross-entropy* (Kullback, 1959) approach have been applied to a wide variety of estimation problems under limited information. A general introduction to *maximum entropy* econometrics, including a large number of extensions and applications can be found in Golan et al. (1996). Further applications include the spatial distribution of commodity flows, passenger trips or agricultural production (see inter alia Wilson, 1967, 1970, 1971; Nijkamp, 1975; You and Wood, 2006; You et al., 2009), problems of efficient aggregation (Batty, 1974), the parameterization of CGE or nonlinear I–O models (Arndt et al., 2001; Fernandez-Vazquez, 2015) and stochastic properties of economic source data (Rodrigues, 2015).

In the context of estimating and updating I–O Tables and Social Accounting Matrices (SAMs), especially the RAS method, which belongs to the *minimal cross-entropy* approaches, is widely used

and has been subject to a number of extensions and generalizations (see Batten, 1983; Golan et al., 1994; Gilchrist and St. Louis, 1999; Robinson et al. 2001; Junius and Oosterhaven, 2003; Huang et al., 2008; Lenzen et al., 2009; Temurshoev et al., 2011; Caggiani et al., 2014; Lenzen et al., 2014; Rodrigues, 2014 as well as the special issue from 2004 of this journal).

However, in all of these applications the values to be estimated, as well as the partial information on which the estimation is based, are either measured in one and the same unit or the problem of different units is circumvented by employing step-wise procedures.

The main innovation of the model developed in this paper is to handle partial information measured in two or more different units in an integrated manner. The key idea is the following: In addition to estimating the flows of aggregated groups of commodities between regions or sectors, the objective is to estimate its composition of individual commodities along with their corresponding prices or caloric contents per ton, such that mass-, financial- and/or energy-balances are simultaneously satisfied. As long as the individual commodities can be assigned to the commodity classifications for which partial information is available, problems of mismatches and different levels of aggregation are resolved at the same time.

Although the model is described in the context of an estimation problem of interregional trade combining economic and transportation data, it offers the flexibility to be applied in any other task, where commodity flows in various units are to be estimated. Most notably, by discussing differences in using physical I–O tables versus environmentally extended monetary I–O tables for environmental impact assessment, Weisz and Duchin (2006) show that discrepancies in the outcomes can be attributed to an assumption implicitly made when monetary I–O data are used: the homogeneity of commodity prices across all uses. Merciai and Heijungs (2014) argue that using monetary I–O tables bears the danger of delivering biased results, due to the violation of mass balances, while Majeau-Bettez et al. (2016) show that imbalances can also be the result of aggregating heterogeneous products consumed in different proportions by the various users. Therefore, the simultaneous estimation model presented here also constitutes a contribution to the reduction of such inconsistencies.

The remainder of this paper is organized as follows. Section 3.2 revisits the classical *maximum entropy* model with an example for a spatial system of commodity flows. In this section we assume that data are measured in a single unit and available in a single classification. Based on this standard model Section 3.3 develops a *maximum entropy* model that is capable of estimating commodity flows measured in different units simultaneously, under the assumption of partial and incomplete data at different levels of aggregation and in mismatching classifications. In section 3.4, a Monte-Carlo analysis is conducted in order to assess the accuracy of estimates of the simultaneous estimation model and to compare its performance with a simple step-wise procedure. Section 3.5 discusses possible extensions of the model and concludes.

3.2 The classical maximum entropy model for estimating commodity flows

As a basis for the model developed in the next section, the principle of *maximum entropy* estimation with its classical application to the estimation of interregional flows developed by Wilson (1971) is recapitulated here.

Let us suppose a set of regions that trade distinguishable commodities $c = 1, \dots, C$, whereby $r = 1, \dots, R$ denotes the suppliers (the region of origin) and $s = 1, \dots, S$ denotes the purchasers (the region of destination). The *micro states* $t_c = (r, s)$ of the system describe the movement of each commodity c from region r to region s. By contrast, the *macro state* of the system counts the number of commodities shipped from region r to region s and can be written as t^{rs}, while $C = \sum_r \sum_s t^{rs}$ represents the total amount of bilateral transactions in the system.

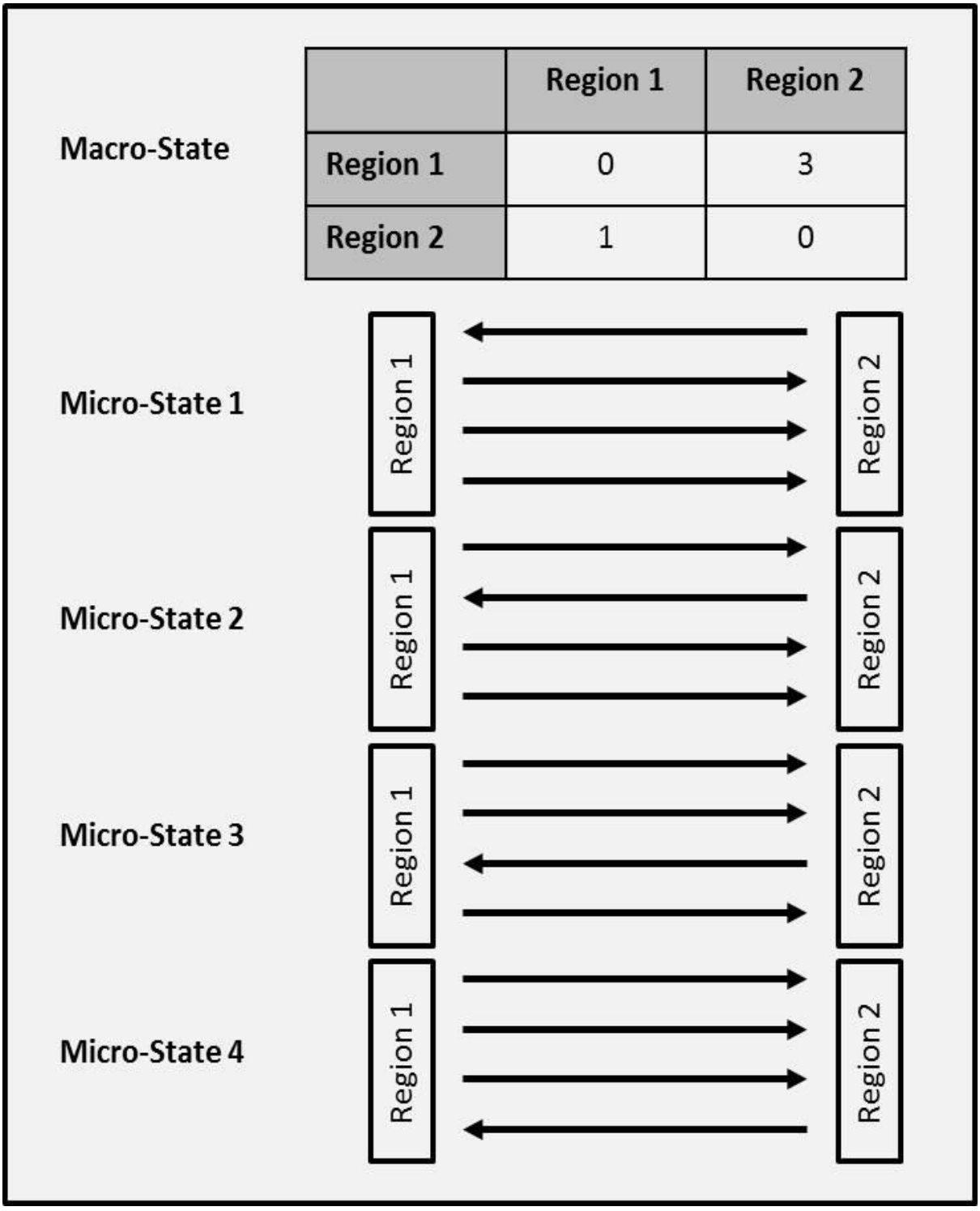
Now, our objective is to estimate the amount of the commodities shipped from r to s. In absence of any information other than the knowledge about the total number of commodities C shipped between all pairs of regions, Jaynes (1957) suggests a logical principle similar to the Laplacian principle of insufficient reason: All possible *micro states* of the system consistent with the partial information about the *macro state* have the same probability, while each *micro state* that is inconsistent with our knowledge about the *macro state* has zero probability. The important consequence of this principle is that the probability to observe any arbitrary *macro state* is proportional to the number of possible *micro states* that yield that *macro state* through aggregation (Snickars and Weibull, 1977). The number of *micro states* that yield a certain *macro state* - i.e., the number of ways in which C units can be distributed into (RS) groups of potential transactions - is given by the combinatorial formula

$$W = \frac{C!}{\prod_r \prod_s t^{rs}!}. \tag{3.1}$$

The most probable distribution of bilateral transactions between the regions (*macro state*) is then found through maximization of Equation 3.1, which yields the *macro state* that can be produced by the maximum number of different *micro states* and is, thus, the most probable one.

A simplified example taken from Sargento (2009) is shown in Figure 3.1, to illustrate these basic ideas. In this example, there are $C = 4$ commodities traded between $R = S = 2$ regions. In addition, we know that $t^{1\,2} = 3$ units of commodities are shipped from region 1 to region 2 and $t^{2\,1} = 1$ unit is shipped from region 2 to region 1. Applying Equation 3.1 to this example shows that this *macro state* can be generated by $4!/(3!\,1!) = 4$ different *micro states*, since each of the four commodities could be the one that is sold by region 2 to region 1. The *macro state* that can be produced by the maximum number of different micro states (i.e., the solution that maximizes W) is that of an even distribution of the four units available in the system (i.e., $t^{11} = t^{12} = t^{21} = t^{22} = 1$).

Figure 3.1 Illustration of the difference between micro- and macro-state descriptions of commodity flow systems



Source: Own elaboration

Dividing the amount of commodities shipped from region r to region s t^{rs} by the total amount of commodities in the system, C, yields the corresponding expression in terms of frequencies $p^{rs} = t^{rs}/C$, where $\sum_r \sum_s p^{rs} = 1$. Taking logs of Equation 3.1 and using Stirling's Approximation[3] delivers Shannon's (1948) *entropy* measure of a discrete probability distribution written in matrix notation

$$H(\boldsymbol{p}) = -\boldsymbol{p}' \ln \boldsymbol{p} \tag{3.2}$$

where $\boldsymbol{p}$ is a column-vector of length (RS) whose generic element p^{rs} denotes the fraction of the commodities totally available in the system that is sold from r to s.

The *entropy* of a distribution is an inverse measure of the degree of information and reaches its maximum for $p^{11} = \cdots = p^{1s} = p^{2s} = \cdots = p^{rs}$. Under the principle of *maximum entropy* Jaynes argues that "in making inferences on the basis of partial information we must use that probability distribution which has the *maximum entropy* subject to whatever we know. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information, which by hypothesis we do not have" (Jaynes, 1957, p. 623).

For the utilization of the principle of *entropy maximization*, additional information about the *macro states* may be added in terms of constraints on the *entropy maximizing* probability distribution $max\, H(\boldsymbol{p})$. In the case of the estimation of commodity flows, this in particular concerns information about row and column totals, such as total tons loaded and unloaded in a region, total regional supply and demand or, in the case of intersectoral flows, total intermediate sales and purchases by sector. Generally, K available data points can be arranged as a stacked column vector $\boldsymbol{k}$. In the case of a doubly constrained model of spatial commodity flows (see Wilson, 1967; Nijkamp, 1975), for example, $\boldsymbol{k}$ is of length $K = R + S$ and contains the partial information about the *macro state* in terms of the S column totals, $\bar{p}^s = \sum_r \bar{t}^{rs}/\bar{C}$, and R row totals $\bar{p}^r = \sum_s \bar{t}^{rs}/\bar{C}$, where the bar indicates available data points.[4] The *entropy maximizing* distribution subject to our information on row and column totals is found by solving the nonlinear program

$$max\, H(\boldsymbol{p}) = -\boldsymbol{p}' \ln \boldsymbol{p} \tag{3.3a}$$

s.t.

$$\boldsymbol{k} = \boldsymbol{G}_{rs}\, \boldsymbol{p} \tag{3.3b}$$

$$1 = \boldsymbol{i}'_{RS} \boldsymbol{p} \tag{3.3c}$$

and subject to the non-negativity constraint

$$\boldsymbol{p} \geq \boldsymbol{0} \tag{3.3d}$$

[3] $\ln x! \cong x \ln x - x$, see Wilson (1967).

[4] Note, that in doubly constraint settings $\boldsymbol{k}$ is actually of length $K - 1$, as the $(R + S)^{\text{th}}$ constraint follows from the other $K - 1$ constraints and ,thus, becomes redundant.

where $\boldsymbol{G_{rs}}$ is a concordance matrix of dimension $(R+S) \times (RS)$, whose elements take the values of 0 or 1 and which relates the frequencies we would like to estimate with the available information in the form of adding up constraints. Hence, in this notation, Constraint 3.3b summarizes the more common notation of a double-constraint, where $\bar{p}^s = \sum_r p^{rs}$ and $\bar{p}^r = \sum_s p^{rs}$ represent consistency requirements on the target matrix with respect to known row and column totals. $\boldsymbol{i}'_{RS}$ is a unit vector of length (RS) used for summation.

The solution of (3.3) is found by solving the first order conditions of the corresponding Lagrangian

$$L(\tilde{\boldsymbol{p}}) = -\tilde{\boldsymbol{p}}' \ln \tilde{\boldsymbol{p}} + (\boldsymbol{k} - \boldsymbol{G_{rs}}\,\tilde{\boldsymbol{p}})'\tilde{\boldsymbol{\lambda}} + (1 - \boldsymbol{i}'_{rs}\tilde{\boldsymbol{p}})'\tilde{\mu}. \tag{3.4}$$

where $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\mu}$ are Lagrangian multipliers and '∼' indicates estimates. The first order conditions are then

$$\frac{\partial L}{\partial \boldsymbol{p}} = -\ln \tilde{\boldsymbol{p}} - \boldsymbol{1} - \boldsymbol{G_{rs}}'\tilde{\boldsymbol{\lambda}} - \tilde{\mu} = \boldsymbol{0} \tag{3.5a}$$

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = \boldsymbol{k}' - \boldsymbol{G_{rs}}\tilde{\boldsymbol{p}} = \boldsymbol{0} \tag{3.5b}$$

$$\frac{\partial L}{\partial \mu} = 1 - \boldsymbol{i}'_{RS}\tilde{\boldsymbol{p}} = 0 \tag{3.5c}$$

Solving system (3.5), with $(RS) + (R+S) + 1$ equations, for $\tilde{\boldsymbol{p}}$ in terms of $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\mu}$ delivers the solution

$$\tilde{\boldsymbol{p}} = \Omega\big(\tilde{\boldsymbol{\lambda}}\big)^{-\boldsymbol{1}} exp\big(-\boldsymbol{G_{rs}}'\tilde{\boldsymbol{\lambda}}\big), \tag{3.6}$$

where the normalization factor $\Omega\big(\tilde{\boldsymbol{\lambda}}\big) = \boldsymbol{i}'_{rs} exp\big(-\boldsymbol{G_{rs}}'\tilde{\boldsymbol{\lambda}}\big)$ is used to convert relative probabilities into absolute ones. The Lagrangian multipliers $\tilde{\boldsymbol{\lambda}}$ reflect the relative contribution of each data point to the optimal value of the objective function and can, hence, be interpreted as a measure of the information content of each data point (Golan et al., 1996). Since there are $(RS) + (R+S) - 1$ independent equations, but $(RS) + (R+S)$ unknowns, an analytical solution of (3.5) can only be made up to a scalar.

If prior information on the flows is available, it can be used for the estimation by modifying the objective. Instead of maximizing the *entropy* of unknown fractions of flows, the *cross-entropy*, i.e., the *entropy* distance also known as *Kullback-Leibler divergence* (Kullback and Leibler, 1951; Kullback, 1959), between the target and the prior matrix is minimized. The principle of *minimal cross-entropy* is also known as the principle of *minimal information gain*, as the information gained when using the target instead of the prior distribution is minimized. Such prior information could, for example, be a table of a previous year or one constructed on the basis of non-survey methods (see Flegg et al., 1995; Miller and Blair, 2009; Többen and Kronenberg, 2015). Therefore, the *maximum entropy* problem can be considered as a special case of the more *general minimal* cross-*entropy* problem where priors are evenly distributed, i.e., in the absence of prior information (Golan et al., 1996).

In the doubly constrained maximum entropy or minimal cross-entropy models, the probability of a certain flow $\tilde{p}^{rs}$ depends on the Lagrangian multipliers of the sending and of the receiving region or sector, $\tilde{\lambda}^{r}$ and $\tilde{\lambda}^{s}$, which can be given an economic interpretation: For example, as *push*- and *pull-effects* in the context of interregional spillovers (Nijkamp, 1975), or as *fabrication*- and *substitution effects* in the case of interindustry transactions (Miller and Blair, 2009).

Depending on the task, additional constraints can be added. In interregional trade applications, for example, additional restrictions on trade or transportation costs can be used to integrate information regarding the spatial dimension of flows. Wilson (1967) shows that the gravity trade model, originally suggested by Leontief and Strout (1963), can be derived from the doubly-constraint *maximum entropy* model with an additional transportation cost constraint. In the optimal solution, the flows between two regions are proportional to the level of supply of r and to the level of demand of s, and proportional to the reciprocal of the costs of trade between both regions. Batten (1983) estimates subnational MRIO tables with a *maximum entropy* model using accounting balances as well as data such as national I–O tables, transportation costs and regional aggregates as constraints.

3.3 Estimating physical and monetary commodity flows simultaneously

In the previous section, the units of measurement did not play any role for the formulation of the estimation problem, as it was implicitly assumed that both the values to be estimated and the row and column totals are measured in the same units. Furthermore, no distinction has been made between different types of commodities. For situations, however, where we want to estimate flows of a variety of different types of commodities on the basis of data in different product classifications and measured in different units, the above simple *entropy* model requires a number of extensions.

3.3.1 A commodity flow system in mixed units of measurement

Assume that we want to estimate spatial flows of different types of commodities under information about row and column sums measured in physical and monetary units, e.g., the amount of tons loaded and unloaded in a region from freight transportation data measured in tons and regional data of supply and demand measured in currency units. In addition it is assumed that row and column totals are available at different levels of aggregation and in mismatching commodity classifications.

The row and column totals measured in tons are available for H different categories of commodities, with, $h = 1, \dots, H$ being a set of commodity groups corresponding to the classification used for freight transportation data. By contrast the row and column totals measured in currency units are available for I different categories, with $i = 1, \dots, I$ denoting commodity groups of the classification used for economic data. Further, it is assumed that $H \neq I$. Classification mismatches occur in this context, if one classification cannot be derived from the other by means of simple aggregation.

The set of row and column totals measured in physical units can then be posed as, respectively,

$$\bar{t}_h^r = \sum_s t_h^{rs} \; \forall \, r = 1, \dots, R; h = 1, \dots, H \text{ and } \bar{t}_h^s = \sum_r t_h^{rs} \; \forall \, s = 1, \dots, S; h = 1, \dots, H, \tag{3.7}$$

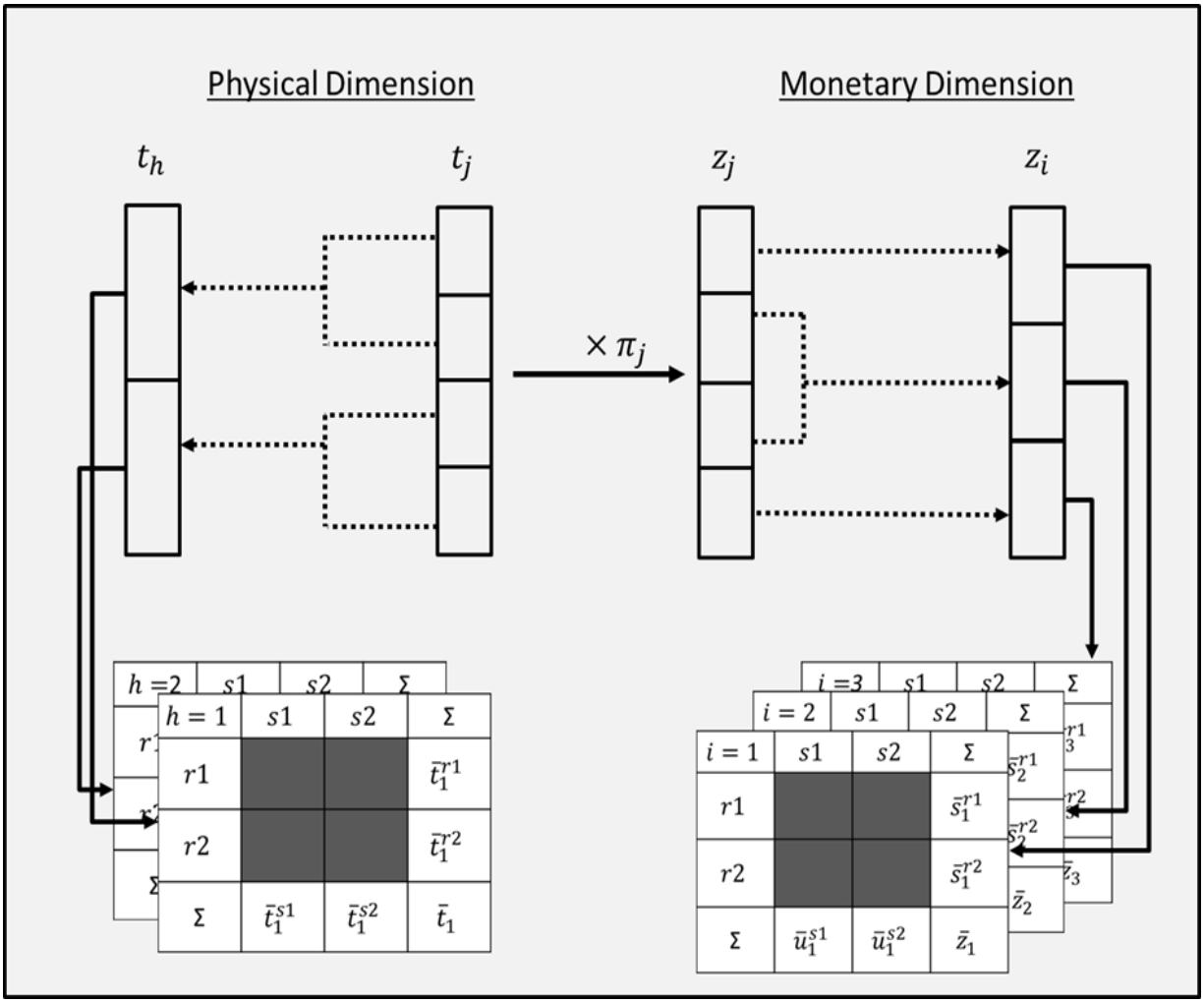
where $\bar{t}_h^r$ are the total shipments of commodity h within and out of region r and $\bar{t}_h^s$ are the total shipments within and to region s, both measured in tons.

The set of row and column totals measured in currency units can be written as, respectively,

$$\bar{z}_i^r = \sum_s z_i^{rs} \; \forall \, r = 1, \dots, R; i = 1, \dots, I \text{ and } \bar{z}_i^s = \sum_r z_i^{rs} \; \forall \, s = 1, \dots, S; i = 1, \dots, I, \tag{3.8}$$

where z_i^{rs} denotes the monetary value of the sales of commodities of type i from region r to region s, $\bar{z}_i^r$ denotes total supply of i by region r (excluding sales to outside regions) and $\bar{z}_i^s$ denotes total demand for commodities of type i by region s (excluding purchases from outside regions).

Figure 3.2 Illustration of the relationship between physical and monetary commodity flows and the alignment of both dimensions through the auxiliary root classification



Source: Own elaboration

In order to align data from both sources, an auxiliary classification $j = 1, \dots, J$ from which both classifications of our data can be derived through simple aggregation needs to be defined. Note that this approach is similar to the 'root-classification' used in the Australian IELab (see Lenzen et al., 2014). The root classification j has to be constructed in such a way that each commodity of type j belongs to exactly one commodity group of classification h and at the same time belongs to exactly one commodity group of classification i.

The relationship between the three commodity-classifications and the general approach of linking data in different units of measurement and in different classifications is illustrated in Figure 3.2. In this example, row and column totals measured in physical units are available for two different types of commodities, while for the monetary dimension row and column totals comprise three commodity groups. Both classifications are mismatching, because both commodity groups of h comprise commodities that belong to two different groups of i: $h = 1$ consists of commodities belonging to $i = 1,2$ and $h = 2$ consists of commodities belonging to $i = 2,3$. The mismatch is resolved through the introduction of the root classification j from which the other classifications can be derived through aggregation.

Formally, the relationship of the root classification to the classifications in which data are available can be expressed in terms of the concordance matrices $\boldsymbol{G_{ij}}$ and $\boldsymbol{G_{hj}}$, whose elements g_{ij} and g_{hj} are equal to one if a commodity j belongs to the commodity groups i or h, respectively, and are equal to zero otherwise.

By making use of the root classification j, the consistency restrictions for the physical flows (3.7) can be rewritten as

$$\bar{t}_h^r = \sum_s \sum_j g_{hj} t_j^{rs} \; \forall \, r = 1, \dots, R; h = 1, \dots, H \tag{3.9a}$$

and

$$\bar{t}_h^s = \sum_r \sum_j g_{hj} t_j^{rs} \; \forall \, s = 1, \dots, S; h = 1, \dots, H. \tag{3.9b}$$

where t_j^{rs} denotes the amount of tons of commodities of the root classification j shipped from r to s.

For expressing the consistency constraints for the monetary side in terms of the root classification, we utilize the property that the monetary value of transactions z_i^{rs} can be expressed as the sum of the amount of tons of j belonging to i times the respective (unknown) price per ton π_j^{rs} of that transaction. Therefore, (3.8) can be rewritten as

$$\bar{z}_i^r = \sum_s \sum_j g_{ij} \pi_j^{rs} t_j^{rs} \; \forall \, r = 1, \dots, R; i = 1, \dots, I \tag{3.10a}$$

and

$$\bar{z}_i^s = \sum_r \sum_j g_{ij} \pi_j^{rs} t_j^{rs} \; \forall \, s = 1, \dots, S; i = 1, \dots, I. \tag{3.10b}$$

Consequently our estimation problem here does not only deal with the flow of commodities from one region (or sector) to another, but needs to be extended to, firstly, the estimation of the composition of aggregate flows with distinct commodities and, secondly, the estimation of their respective prices per ton. Hence, the objective is to find a distribution of between RS pairs of regions (or sectors), a composition of these flows of distinct commodities j and the corresponding prices of these commodities that are optimal in terms of the *maximum entropy principle*.

3.3.2 Entropy measures for quantities and prices

For aggregate commodity flows the *entropy* that measures the heterogeneity of flows between RS pairs of regions has been introduced by means of Equation 3.2 in the previous section. In this section, the *entropy* measure for our commodity flow system has to be extended into two directions: First, an *entropy* measure for describing the heterogeneity of the composition of aggregate flows in terms of the root classification j is required. Second, we need an *entropy* measure describing the uncertainty of average prices at which commodities of type j are traded.

The commodity composition of the interregional flows

Let p_j^{rs} be the fraction of commodity j in the fraction of total flows shipped between r and s, p^{rs}, such that $\sum_j p_j^{rs} = p^{rs} = t^{rs}/C$. Following Theil (1966), the *entropy* of the internal composition of flow t^{rs} can be measured by

$$H_{rs}\left(p_j^{rs}\right) = -\sum_j \frac{p_j^{rs}}{p^{rs}} \ln \frac{p_j^{rs}}{p^{rs}} \tag{3.11}$$

Equation 3.11 is equivalent to Theil's (1966) conception of the *within-set entropy*, which he uses to measure the internal heterogeneity of sets, e.g., in cases where observations have been aggregated to groups. Opposed to that, the *between-set entropy* is represented by the *entropy* measure of Equation 3.2.

Both, the *within-set* and the *between-set entropy* are connected through the following relationship that describes the total *entropy* in the commodity flow system

$$H_p\left(p^{rs}, p_j^{rs}\right) = H(p^{rs}) + \sum_r \sum_s p^{rs} H_{rs}\left(p_j^{rs}\right) \tag{3.12a}$$

or

$$H_p\left(p^{rs}, p_j^{rs}\right) = -\sum_r \sum_s p^{rs} \ln p^{rs} - \sum_r \sum_s p^{rs} \sum_j \frac{p_j^{rs}}{p^{rs}} \ln \frac{p_j^{rs}}{p^{rs}}. \tag{3.12b}$$

where the first term on the ride-hand side measures the *between-set entropy* of aggregate flows between each pair of regions and the second term measures the *within-set entropy* of each aggregate flow.

The total *entropy* of this commodity flow system can also be expressed in terms of p_j^{rs} only. For a more convenient expression substitute $\sum_j p_j^{rs} = p^{rs}$ into Equation 3.12b, which yields

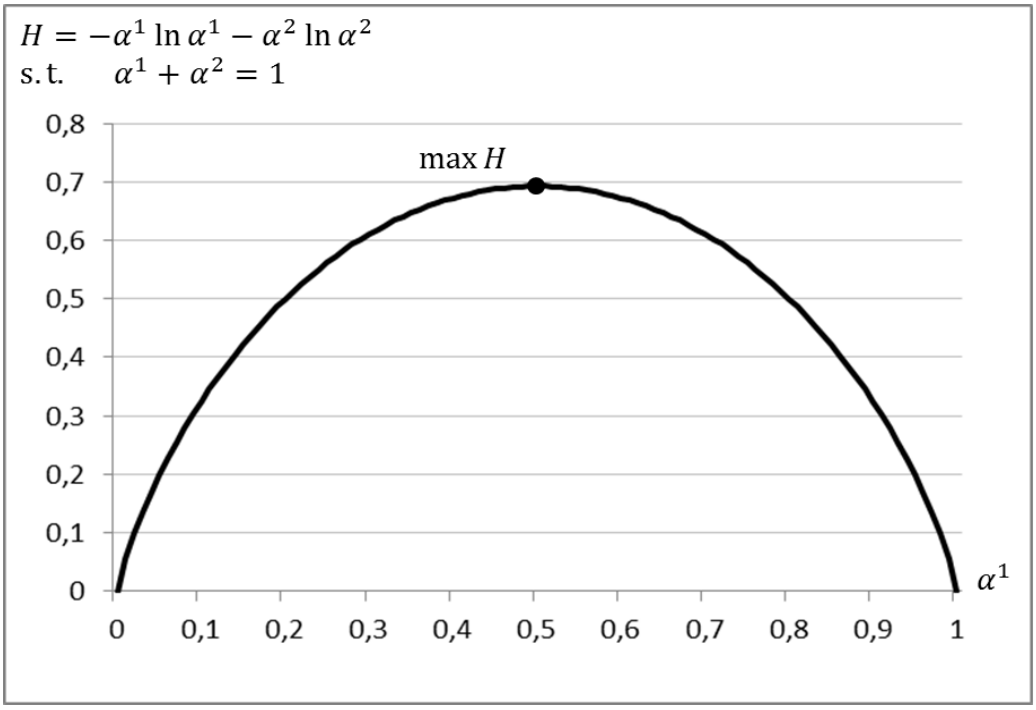
$$H_p\left(p_j^{rs}\right) = -\sum_r \sum_s \sum_j p_j^{rs} \ln p_j^{rs}. \tag{3.13}$$

The uncertainty of prices

In addition to the unknown fractions of commodity flows, p_j^{rs}, we need to estimate their average prices per ton. Typically, information about prices may be gained through production or trade statistics from databases such as UN-COMTRADE or PRODCOM, which publish annual import, export and production data in tons and currency at high commodity resolution for virtually all countries in the world. It is important to note that the prices have to be understood as an unknown weighted average price of unobserved transactions of commodities of type j from r to s. Even at the highest resolution at which commodity specific data are usually published, each group comprises of a large variety of different types, variants and brands. In addition, even prices for one and the same commodity will be different due to heterogeneous seller-buyer relationships.

Information from such databases can be used to construct so-called *supports* for the estimation of average prices of a commodity of type j. *Supports* can be upper or lower bounds, observations, distribution parameters or any other knowledge (or beliefs) that describe the distribution of prices within commodity group j (see Golan, 1996).

Figure 3.3 Shape of the entropy measure for two supports



Source: Own elaboration

Let us assume that the only information about prices we have are observations of the maximal and the minimal prices per ton at which commodities of type j are traded (e.g., from export statistics), $\bar{\pi}_j^{max}$ and $\bar{\pi}_j^{min}\ \forall\, r,s$, respectively. If we assume that a plausible range of the unknown price, π_j^{rs}, falls in the interval of $\left[\bar{\pi}_j^{min}, \bar{\pi}_j^{max}\right]$, then π_j^{rs} can be expressed as a linear combination of its bounds and their respective weights $\propto_j^{rs1}$ and $\propto_j^{rs2}$:

$$\pi_j^{rs} = \propto_j^{rs1} \bar{\pi}_j^{max} + \bar{\pi}_j^{min} \propto_j^{rs2}, \tag{3.14}$$

where $\propto_j^{rs1}, \propto_j^{rs2} \geq 0$ and $\propto_j^{rs1} + \propto_j^{rs2} = 1$ and *supports* are the lower and upper bounds $\left[\bar{\pi}_j^{min}, \bar{\pi}_j^{max}\right]$. The estimates of the unknown prices are, then, gained through maximization of the *entropy* of their respective weights. Figure 3.3 shows the *maximum entropy* for different values of α^1 between zero and one, subject to $\alpha^1 + \alpha^2 = 1$. It can be seen that *entropy* of α^1 and α^2 has the shape of an inverse U and takes its maximum for $\alpha^1 = \alpha^2 = 0.5$. Thus, in absence of any other constraints, price estimates resulting from the maximization of this *entropy measure* will be the arithmetic mean of the upper and the lower bound.

In the more general case of *M supports*, the unknown average prices per ton, π_j^{rs}, can be expressed as a convex combination of *M supports* $\boldsymbol{\bar{\pi}_j} = \left[\bar{\pi}_j^1, \dots, \bar{\pi}_j^m, \dots, \bar{\pi}_j^M\right]$ and *M* weights $\boldsymbol{\alpha_j^{rs}} = [\alpha_j^{rs1}, \dots, \alpha_j^{rsm}, \dots, \alpha_j^{rsM}]$ that sum up to one (see Golan, 1996; Chapter 6). The corresponding *entropy measure* of prices may, then, be written as

$$H_\alpha\left(\alpha_j^{rsm}\right) = -\sum_r \sum_s \sum_j \sum_m \alpha_j^{rsm} \ln \alpha_j^{rsm} \tag{3.15}$$

3.3.3 The combined estimation model

When combining the *entropy measures* for the unknown fractions of commodity flows and the unknown weights on *price-supports*, an important aspect to consider are the weightings of both objectives relative to each other. If no weights are explicitly assigned, implicit weightings are made based on the *maximum entropies* of the objectives in the unconstrained case, i.e., when all unknowns are evenly distributed.

For example, if we want to estimate $RSJ = 1000$ unknown commodity flows and prices, whereby two *price-supports* are available for each bilateral transaction of j, then the *entropy measure* for the commodity flows approaches its maximum for $p_1^{11} = \dots = p_j^{rs} = 1/1000\ \forall r,s,j$. In the case of the unknown weights on the *price-supports*, however, the *entropy measure* approaches its maximum for $\alpha_j^{rs1} = \alpha_j^{rs2} = 0.5\ \forall rsj$. While the unknown fractions of flows sum up to one, the unknown weights on the *price-supports* sum up to 1000. As a consequence, substituting $p_1^{11} = \dots = p_j^{rs} = 1/1000$ into (3.13) yields 6.91, while substituting $\alpha_j^{rs1} = \alpha_j^{rs2} = 0.5$ into (3.15) yields 693. Hence, in such an

estimation problem, the objective for the estimation of prices receives roughly a hundred times as much weight as the estimation of the fractions of flows. For this reason, if one wants to assign equal weight on both objectives, each of them needs to be divided by their respective unconstrained maximum entropy.

In general, the unconstrained *maximum entropies* in a system with RSJ bilateral transactions and M *price-supports* for each transaction are $\ln RSJ$ and $RSJ \ln M$, respectively. Dividing the *entropy measures* (3.13) and (3.15) by their unconstrained *maximum entropies* delivers their respective *relative entropies* (see Golan et al., 1996):

$$H_p^*(p_j^{rs}) = -\frac{\sum_r \sum_s \sum_j p_j^{rs} \ln p_j^{rs}}{\ln RSJ} \tag{3.16a}$$

and

$$H_\alpha^*(\alpha_j^{rsm}) = -\frac{\sum_r \sum_s \sum_j \sum_m \alpha_j^{rsm} \ln \alpha_j^{rsm}}{RSJ \ln M}. \tag{3.16b}$$

Based on these *relative entropy* measures and on the data consistency constraints shown in Equations 3.9a, 3.9b, 3.10a and 3.10b, the *entropy maximization* problem can, then, be posed as

$$\max H^*(\boldsymbol{p}, \boldsymbol{\alpha}) = H_p^*(\boldsymbol{p}) + H_\alpha^*(\boldsymbol{\alpha}) = -\frac{\boldsymbol{p}' \ln \boldsymbol{p}}{\ln RSJ} - \frac{\boldsymbol{\alpha}' \ln \boldsymbol{\alpha}}{RSJ \ln M} \tag{3.17a}$$

subject to the data consistency constraints of the physical dimension

$$\boldsymbol{k_h} = \left(\boldsymbol{G_{rs}} \otimes \boldsymbol{G_{hj}}\right)\boldsymbol{p} \tag{3.17b}$$

and the re-parameterized constraints of the monetary dimension, which are gained through substituting $\pi = \boldsymbol{\alpha}' \overline{\boldsymbol{\Pi}}$ into the Equations 3.10a and 3.10b

$$\boldsymbol{k_i} = \left(\boldsymbol{G_{rs}} \otimes \boldsymbol{G_{ij}}\right)\boldsymbol{\alpha}' \overline{\boldsymbol{\Pi}} \boldsymbol{p}, \tag{3.17c}$$

subject to the adding-up conditions

$$1 = \boldsymbol{i}_{jrs}' \, \boldsymbol{p} \tag{3.17d}$$

and

$$\boldsymbol{i}_{jrs} = \boldsymbol{I_{jrs}} \otimes \boldsymbol{i}_m' \boldsymbol{\alpha}, \tag{3.17e}$$

as well as to the non-negativity constraint

$$\boldsymbol{p} \geq \boldsymbol{0} \tag{3.17f}$$

$$\boldsymbol{\alpha} \geq \boldsymbol{0}, \tag{3.17g}$$

where $\boldsymbol{p} = \left[\boldsymbol{p^{11}}, \dots, \boldsymbol{p^{rs}}, \dots, \boldsymbol{p^{RS}}\right]'$ with $\boldsymbol{p^{rs}} = \left[p_1^{rs}, \dots, p_J^{rs}\right]'$ is a vector length JRS of unknown fractions of tons and $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha_1^{rs}}, \dots, \boldsymbol{\alpha_j^{rs}}, \dots, \boldsymbol{\alpha_J^{rs}}\right]$ with $\boldsymbol{\alpha_j^{rs}} = \left[\alpha_j^{rs1}, \dots, \alpha_j^{rsM}\right]$ is a vector of length $MJRS$

of unknown weights on the *price-supports* $\bar{\boldsymbol{\Pi}}$. The JRS unknown fractions of tons are connected to $H(R+S)$ physical row and column totals, $\boldsymbol{k_h}$, via the concordance matrix $\boldsymbol{G_{rs}} \otimes \boldsymbol{G_{hj}}$, where $\otimes$ is the Kronecker product. As in the previous section, the physical row and column total are scaled by the total amount of commodities available, $\bar{C}$. Therefore, the generic elements of $\boldsymbol{k_h}$ are $\bar{p}_h^s = \bar{t}_h^s/\bar{C}$ and $\bar{p}_h^r = \bar{t}_h^r/\bar{C}$ and denote the shares of commodities belonging to group h in the total amount of commodities $\bar{C}$ available in the system that are delivered to s or from r, respectively.

The monetary row and column totals are scaled with $\bar{C}$, as well, such that the generic elements of $\boldsymbol{k_i}$ are $\bar{\pi}_i^s = \bar{z}_i^s/\bar{C}$ and $\bar{\pi}_i^r = \bar{z}_i^r/\bar{C}$. As a consequence, they are transformed into constraints on the weighted average prices $p_j^{rs}\pi_j^{rs} = p_j^{rs}\sum_m \alpha_j^{rsm}\bar{\pi}_j^m$, where the estimated fractions, p_j^{rs}, represent the weights. The estimates of the weighted average prices are, then, mapped on the corresponding $I(R+S)$ row and column totals via $\boldsymbol{G_{rs}} \otimes \boldsymbol{G_{ij}}$.

3.4 Monte-Carlo Simulation

In this section, the performance of the *maximum entropy* model developed above is assessed by means of a Monte Carlo simulation. For this purpose, we generate 500 random benchmark setups. Each of these setups, consist of $RSJ = 400$ random commodity flows measured in tons and corresponding random prices per ton. Afterwards, we estimate these from the "known" aggregate information about commodity flows and prices. In addition to assessing the quality of estimates gained from the simultaneous estimation model, we also compare its performance against a simplified step-wise procedure.

In each benchmark setup, we distinguish $R = S = 10$ regions and $J = 4$ commodity groups. The $J = 4$ commodity groups are assumed to add up to two classifications: viz. $H = 2$ for the physical and $I = 3$ for the monetary dimension. The adding up rules and corresponding concordance matrices are taken from the example for mismatching classifications shown in Figure 3.2.

In order to get benchmark setups that are close to a real commodity flow system, the spatial structure of aggregate commodity flows is taken from an origin-destination matrix depicting shipments of machinery products within and between ten out of Germany's 16 federal states. This spatial structure does not vary across the 500 benchmark setups. What varies in each benchmark setup are, firstly, the shares of the $J = 4$ commodity groups in each of the $RS = 100$ aggregate flows and, secondly, the average prices of per ton of each of the $RSJ = 400$ commodity flows. In the next subsection, we describe how these two are randomly drawn.

Specification of the simultaneous estimation model

The 500 benchmark setups we are estimated by means of (3.17), with a slight modification. As we want to focus on assessing the quality of estimates for monetary flows, it is assumed that the origin-

destination matrices measured in tons are given for both commodity groups of classification h, i.e., $\bar{t}_h^{rs}$. As a consequence, the data consistency constraints (3.17b) for the physical dimension are replaced by

$$\bar{t}_h^{rs} = \sum_j g_{hj} t_j^{rs} \quad \forall\, (r,s) = \{(1,1), \dots, (1,S), \dots, (R,S)\}; h = 1,2 \tag{3.17b'}$$

For the monetary dimension, by contrast, row and column totals for the three commodity groups of i are assumed to be known.

Thus, we resemble a typical information context, when monetary trade flows for a subnational MRIO are to be estimated from transportation data measured in tons and published in different levels of aggregation and mismatching classifications. Often these origin-destination matrices are incomplete, in addition. However, in order to keep the setting as simple as possible, we assume that complete origin-destination matrices are available.

With respect to the *price-supports*, we assume three different scenarios for the availability of information. In *Scenario 1*, it is assumed that only the average price of j at the national level can be observed and that the unknown prices may be up to 50% larger or smaller than that average. By contrast, in *Scenario 2* it is assumed that only the minimal and maximal prices for j, $\bar{\pi}_j^{min}$ and $\bar{\pi}_j^{max}$, in each of the 500 benchmark setups can be observed. Finally, in order to assess to what extend the results depend on the quality of the information on prices, it is assumed in *Scenario 3* that the respective price of each flow of j from r to s is perfectly known.

Specification of the step-wise estimation model

The simultaneous estimation model described above is compared against a simplified step-wise estimation model; in order to verify to what extend improvements in the quality of estimates can be gained. Since the classifications of h and i are mismatching, it is not possible to use the full information provided by the two origin-destination matrices for h. Instead, we use an approach similar to that used in Thissen et al. (2013) and compute import- and export coefficients from the aggregates transported from r to s as:

$$EX^{rs} = \bar{t}^{rs} / \bar{t}^{r\cdot} \tag{3.18a}$$

and

$$IM^{rs} = \bar{t}^{rs} / \bar{t}^{\cdot s}\ . \tag{3.18b}$$

The export coefficients are then multiplied with monetary row totals (i.e., total supply of i by region r measured in currency) and import coefficients are multiplied with monetary column totals (i.e., total demand for i by region s measured in currency). Thus, we assume that all commodities traded between r and s have the same average import and export propensities. In this way, we avoid posing

explicit assumptions for the transition between the classifications h and i. Finally, priors of the monetary trade flows for each commodity group i are computed as

$$z_i^{rs\,0} = (EX^{rs}\bar{z}_i^r + IM^{rs}\bar{z}_i^r)/2. \tag{3.19}$$

In the final step, these priors are adjusted to the monetary row and column totals for each commodity group i using RAS.

3.4.1 Drawing random quantities and prices

In order to ensure that flows and prices are of a realistic magnitude, they are randomly drawn from the distribution of tons and prices observed in German export data of machinery products from 2008 at 8-diggit level. The disaggregated commodity groups reported in the export data are, at first, assigned to $J = 4$ categories distinguished in our setup. Thereafter, for each category j we compute means $\boldsymbol{\mu_j} = [\mu_{t_j}, \mu_{\pi_j}]$ and (sample) variances $\boldsymbol{\sigma_j^2} = [{\sigma_{t_j}}^2, {\sigma_{\pi_j}}^2]$ of quantities and prices, as well as the corresponding covariance $\sigma_{t_j \pi_j}$ from the logarithmic values of the export data. These parameters are, then, used to define the joint (i.e., bivariate) distributions of quantities and prices from which the quantities and prices for the benchmark setups are randomly drawn. The distribution parameters, computed from the export data are shown in the table at the top of

Table 3.1.

For each commodity flow j between r and s in each of 500 benchmark setups, we draw n times from the respective bivariate normal distribution of (logarithmic) quantities and prices defined by $\boldsymbol{\mu_j}$ and $\boldsymbol{\Sigma_j}$ using the Matlab's multivariate normal random numbers generator. Here, $\boldsymbol{\Sigma_j}$ is the 2×2 covariance matrix of commodity group j, which contains the respective variances of quantities and prices on the diagonal and their covariance on the off-diagonal. After applying the exponential transformation to the logarithmic values, the amounts of tons of each commodity flow j between r and s, then, results from summing over the n draws. Afterwards, these are adjusted, such that the total amount of tons shipped between r and s from aggregate origin-destination matrix for machinery products are met. The corresponding prices per ton are computed as (ton-) weighted averages over the n draws.

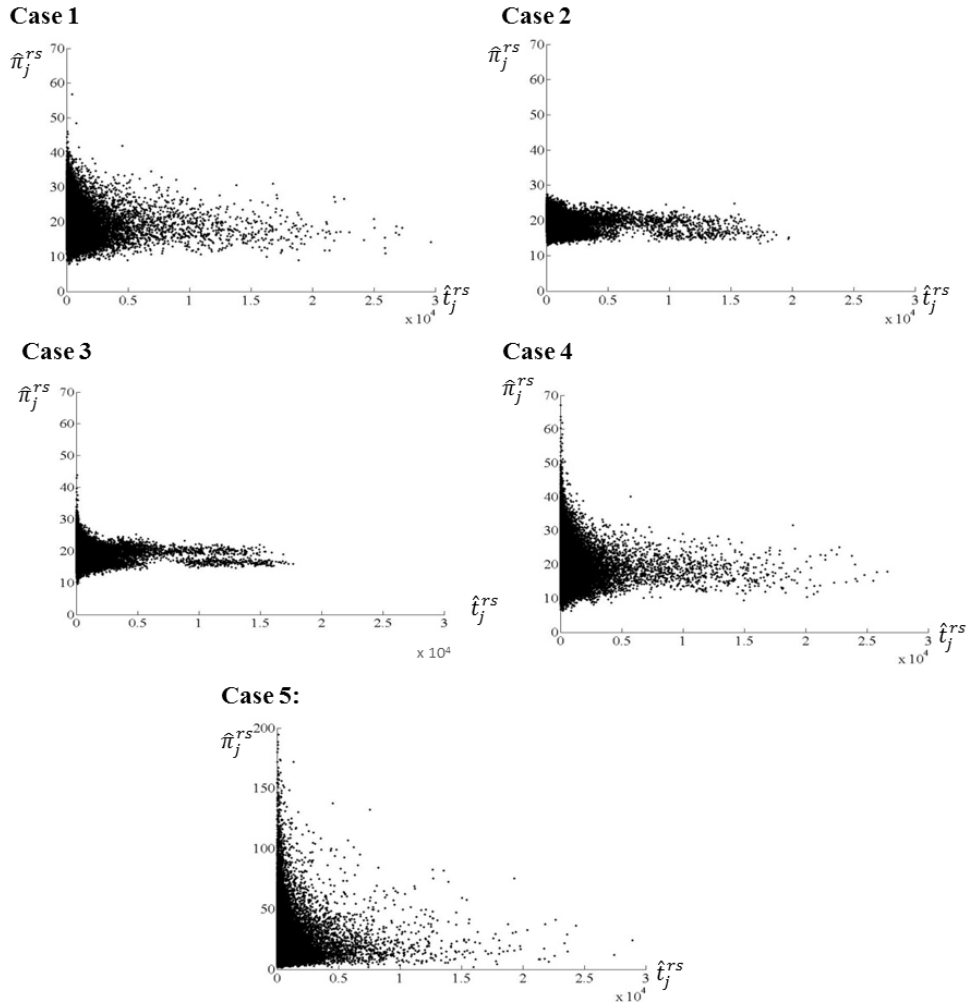
The reason, why we draw n times, is that we want to mimic the characteristic of aggregated commodity groups as being comprised of many different commodities at various prices in different proportions. Due to the law of large numbers, the number of draws determines the average deviation of tons and prices from their respective means and, thus, the degree of uncertainty in each benchmark system. Therefore, we use the number of draws to create five different cases with 100 benchmark setups each, in order to assess the impact of varying degrees of uncertainty in the benchmark setups on the outcomes of the estimation.

Table 3.1 Parameters of the joint distributions of quantities and prices of the commodity groups ***j*** computed from German export data of machinery products in 2008

j	No. 8-diggit	Tons t_j				Prices π_j				$\sigma_{t_j \pi_j}$
		μ_{t_j}	σ_{t_j}	Min	Max	μ_{π_j}	σ_{π_j}	Min	Max	
1	130	8.455	1.577	3.967	11.513	2.796	0.621	1.378	4.302	-0.235
2	111	7.635	1.496	2.862	10.437	2.975	0.673	1.116	4.449	-0.108
3	45	8.281	2.082	3.718	12.128	2.999	0.612	1.677	4.473	-0.384
4	73	7.450	1.801	2.660	11.027	3.074	0.646	0.011	4.689	-0.271

Source: Own calculations.

Figure 3.4 Distribution of quantities and prices in benchmark setups.



Source: Own calculations.

The choice about the number of draws, n, is explained in the following:

- *Case 1* and *Case 2*: Here, each benchmark flow of j from region r to region s and its respective average price per ton are generated by a constant number of draws. In Case 1 the number of draws is $n = 10$, whereas in Case 2 we use $n = 100$. Comparing the scatter-plots for Case 1 and Case 2 in Figure 3.4, shows the impact of the different values of n. The lower the number of draws, n, is, the larger are the deviations of prices from their respective means. This holds true for quantities as well, but it becomes less apparent due to the scaling with the aggregate origin-destination matrix. As consequence, especially prices, are much more variable in Case 1 ($n = 10$) than in Case 2 ($n = 100$).
- *Case 3* and *Case 4*: Here, it is assumed that the degree of uncertainty increases with the size of aggregate flows between r and s from the origin-destination matrix, $\bar{t}^{rs}$. This assumption can be justified with the argument that relatively large flows are more likely to consist of a larger variety of different products traded at different prices compared to relatively small flows. As a consequence, larger flows are more likely to have average prices per ton that are closer to the mean compared to smaller ones. For Case 3 we set the (rounded) number of draws to $n = \sqrt[2]{\bar{t}^{rs}}\ \forall j$ and we set it to $n = \sqrt[4]{\bar{t}^{rs}}\ \forall j$ for Case 4. Thus, the number of draws varies between $n = 200$ and $n = 13$ in the former case, and between $n = 14$ and $n = 4$ in the latter case. Comparing the scatter plots of Case 3 and Case 4 with those of Case 2 and Case 1 in Figure 3.4, respectively, shows that the degree of uncertainty is smaller for relatively large flows, whereas it is larger for relatively small flows.
- *Case 5*: Quantities and prices are drawn independently from two different distribution (i.e., zero covariance, $\sigma_{t_j \pi_j} = 0$). Here, tons are drawn from a uniform distribution taking values between one and 10,000, whereas prices are generated by a single draw from the univariate lognormal distribution defined by the means and variances of the prices of j. For this reason, the degree of uncertainty in Case 5 is by far the largest compared to the four previous cases, because the benchmark values as can be observed in Figure 3.4.

3.4.2 Monte-Carlo Simulation

In this subsection, the performance of the simultaneous estimation model is assessed and its performance is compared against a simple step-wise approach. The outcomes are compared by means of three matrix-comparison statistics: WAPE (weighted absolute percentage deviation), MAPE (mean absolute percentage deviation) and Theil's U_1 coefficient (see Bonfiglio and Chelli, 2008; Pavia et al., 2009). As an example, the statistics for comparing the estimates of $\tilde{z}_j^{rs}$ with its respective benchmark value z_j^{rs} are computed as follows:

$$WAPE = 100\frac{\sum_r\sum_s\sum_j\left|\tilde{z}_j^{rs}-z_j^{rs}\right|}{\sum_r\sum_s\sum_j z_j^{rs}} \tag{3.20}$$

$$MAPE = 100\sum_r\sum_s\sum_j\frac{\left|\tilde{z}_j^{rs}-z_j^{rs}\right|}{RSJ} \tag{3.21}$$

$$U_1 = 100\frac{\sqrt{\sum_r\sum_s\sum_j\left(\tilde{z}_j^{rs}-z_j^{rs}\right)^2}}{\sqrt{\sum_r\sum_s\sum_j\tilde{z}_j^{rs^2}}+\sqrt{\sum_r\sum_s\sum_j z_j^{rs^2}}}. \tag{3.22}$$

First, the performance of the simultaneous approach in estimating monetary commodity flows at the aggregate level of $I = 3$ commodity groups from $H = 2$ (mismatching) origin-destination matrices measured in tons is discussed and compared to the performance of a simplified stepwise procedure.

Table 3.2 shows the outcomes of the simultaneous approach under the three different scenarios for the availability of information on prices (Scen. 1 – Scen. 3) and those of the step-wise procedure in terms of the matrix comparison statistics (3.20) – (3.22). For each of the five cases, the first column ('total') reports the values of (3.20) – (3.22) across all 100 setups, whereas the second and the third column present the minimal and the maximal outcomes among the 100 setups. Table 3.3 shows the number of benchmark setups in which the respective procedures performed best. Since perfect knowledge of prices is unrealistic in real applications, Scenario 3 is excluded from Table 3.3.

Regarding interregional monetary flows (z_i^{rs}**)**, it can be observed that under the assumption that unknown prices may vary in a bandwidth of $\pm 50\%$ around the observed national average (*Scenario 1*), the simultaneous approach outperforms the stepwise procedure across all comparison statistics in all cases, except of Case 5. Across the 100 benchmark setups of Case 5, the simultaneous approach performs worse in terms of WAPE and U_1, but it performs better in terms of MAPE at the same time. These outcomes for Case 5 can be explained by the fact that large ton-flows with prices out of bounds occur far more often than in the other cases. However, the fact that the minimal values of all three matrix comparison statistics are lower indicates that the simultaneous estimation model may potentially outperform a simplified step-wise approach in this extreme case, too.

Table 3.2 Deviations of estimated monetary flows from benchmark values

		Case 1			Case 2			Case 3			Case 4			Case 5		
		total	**min**	**max**	**total**	**min**	**max**	**total**	**min**	**max**	**total**	**min**	**max**	**total**	**min**	**max**
Scen. 1	**WAPE**	17.16	12.60	22.42	6.49	4.69	9.10	7.85	5.73	10.93	8.37	5.71	11.86	39.99	25.19	64.05
	MAPE	35.76	25.32	52.32	9.78	7.62	11.24	17.75	14.30	23.88	13.93	11.36	16.63	109.30	60.88	225.98
	U1	5.05	3.07	8.47	2.14	1.37	3.71	2.18	1.35	3.36	2.61	1.65	4.89	11.00	4.69	27.20
Scen. 2	**WAPE**	19.90	15.81	25.87	6.11	4.34	8.76	8.72	6.26	17.55	8.50	6.22	12.46	57.42	35.98	78.25
	MAPE	43.04	30.97	64.38	10.76	7.95	21.31	20.93	14.56	37.26	15.66	11.24	25.82	198.35	123.53	481.67
	U1	6.12	3.85	9.66	1.88	1.21	3.26	2.48	1.44	4.74	2.57	1.61	5.39	16.26	5.65	28.63
Scen. 3	**WAPE**	12.71	9.15	20.70	4.52	3.03	6.76	5.81	4.19	7.67	6.23	4.17	8.22	14.43	9.40	22.29
	MAPE	30.80	23.20	45.25	8.52	6.38	11.00	15.46	11.59	22.16	12.20	9.93	15.28	51.04	31.81	81.72
	U1	3.48	2.18	6.97	1.34	0.71	2.66	1.48	1.01	2.28	1.80	1.03	2.72	3.18	1.07	7.60
Step-wise	**WAPE**	20.72	15.60	28.61	7.19	5.07	10.65	8.81	6.33	11.31	9.78	7.05	14.22	38.64	26.37	52.23
	MAPE	40.80	31.07	52.33	10.94	8.88	13.44	20.00	16.79	24.33	15.67	12.49	19.52	124.47	80.04	305.36
	U1	6.07	4.02	10.24	2.28	1.31	3.84	2.32	1.46	3.22	2.98	1.71	5.57	10.04	5.09	19.52

Source: Own calculations.

Compared to *Scenario 1*, if only upper and lower bounds of prices are known (*Scenario 2*), the performance improves slightly in few cases with relatively low variation in prices, i.e., Case 2 in terms of WAPE and U_1, as well as in terms of U_1 in Case 4. However, the performance becomes significantly worse across all comparison statistics, when prices are highly variable, i.e., in Case 1, Case 3 and especially in Case 5. If prices are perfectly known (*Scenario 3*), the performance increases significantly across all cases and comparison statistics. This increase in performance is particularly significant for those cases with high degree variation in prices.

Considering the number of times where the simultaneous or the step-wise approach performed best, respectively, it can be observed from Table 3.3 that the former clearly outperforms the latter in all cases except of Case 5. With an exception of U_1 in Case 3 and Case 4, the simultaneous approach performs better in more than 90 out of 100 setups in each of the first four cases. Even in Case 5, where the comparison statistics indicate a worse performance of the simultaneous approach, it performs better in slightly more case compared to the step-wise approach.

Table 3.3 Number of runs in which the simultaneous and step-wise approach performed best

		Case 1	Case 2	Case 3	Case 4	Case 5
Scen. 1	**WAPE**	6	66	29	45	0
	MAPE	0	30	7	19	0
	U1	3	68	25	54	1
Scen. 2	**WAPE**	92	28	66	48	53
	MAPE	95	69	91	81	74
	U1	89	22	47	32	55
Step-wise	**WAPE**	2	6	5	7	47
	MAPE	5	1	2	0	26
	U1	8	10	28	14	44

Source: Own calculations.

Finally, Table 3.4 shows the deviations of the estimates from benchmark tons and prices for the disaggregated commodity groups j. As the step-wise procedure does not generate estimates of tons and prices for the disaggregated commodity groups j, only outcomes of the simultaneous estimation model are shown. At the more disaggregated level the relative deviations from the benchmark values are much larger compared to the outcomes for the more aggregate flows shown in Table 3.2. Comparing Scenario 1 with Scenario 2 shows that the mean deviations (MAPE) of tons and prices in Scenario 1 are larger than in Scenario 2 across almost all of the cases. The only exceptions can be observed for prices in Case 2. Opposed to that, the *supports* used in Scenario 1 deliver better results in terms of WAPE and U_1 in the Cases 2 to 4, while they perform worse in Case 1 and Case 5 across all three comparison statistics. What makes these outcomes peculiar is that the *supports* in Scenario 1 deliver much better results on the more aggregate level across all cases (see Table 3.2 and Table 3.3). The only plausible explanation for this outcome is that individual tons and prices are estimated less accurate in Scenario 1, due to smaller bandwidth of possible prices, but that at the same time the composition of aggregate flows with products of type j and the proportions of prices are estimated more accurately. As we have seen above, perfect knowledge about prices leads to a significant increase in accuracy. This is also the case on the more disaggregate level, whereby the increase in performance is less significant than on the more aggregate level.

Table 3.4 Deviations of estimated tons and prices from benchmark values

		tons					prices				
		Case 1	**Case 2**	**Case 3**	**Case 4**	**Case 5**	**Case 1**	**Case 2**	**Case 3**	**Case 4**	**Case 5**
Scen. 1	**WAPE**	34.56	11.42	15.88	15.67	75.39	17.33	4.84	7.58	6.80	109.71
	MAPE	42.04	11.82	21.06	16.67	81.76	20.00	5.24	10.50	7.53	182.09
	U1	10.02	3.34	4.38	4.57	24.75	10.96	3.25	6.47	4.61	41.18
Scen. 2	**WAPE**	32.85	15.46	17.68	18.28	57.16	15.23	6.48	7.63	7.99	62.03
	MAPE	39.51	11.41	18.76	15.66	69.64	16.33	5.31	8.90	7.30	75.37
	U1	9.97	6.22	6.08	6.79	15.70	10.33	3.40	5.86	4.64	32.05
Scen. 3	**WAPE**	22.05	7.91	9.94	10.77	28.02	-	-	-	-	-
	MAPE	33.82	9.34	16.09	13.30	55.01	-	-	-	-	-
	U1	5.03	1.96	2.19	2.61	6.23	-	-	-	-	-

Source: Own calculations.

3.5 Discussion and conclusion

The objective of this Chapter is the development of a model that is capable to estimate unobserved physical and monetary commodity flows simultaneously under the assumption of limited information. This objective is reached through a *maximum entropy* formulation, where unknown physical flows are estimated along with corresponding prices for transformation into monetary values, such that joint mass and financial balances are simultaneously satisfied. In addition, the model is capable to overcome typical challenges that arise when data from different sources are combined, including classification mismatches and different levels of aggregation.

Usually, such challenges are addressed through combining different procedures for each task within a stepwise approach. In the case of spatial commodity flows, for example, a stepwise approach would typically include the estimation of flows measured tons and their reconciliation to mass balances, the transformation into monetary values, the aggregation or disaggregation and the dissolution of possible classification mismatches and, finally, a second reconciliation to financial balances. As there are many different approaches for each task, this has the disadvantage that outcomes from different applications are often hardly comparable with each other. Furthermore, combining several steps makes the whole procedure prone to errors and embodies the danger to not fully utilize the information available.

In order to assess the accuracy of its estimates, a Monte-Carlo Simulation is conducted, where we use our model to estimate 500 randomly generated benchmark commodity flow systems from information about physical and monetary aggregates and bounds on prices. Furthermore, we also assess the relative performance of our model in comparison to a simple step-wise procedure. Our results show that the simultaneous approach performs significantly better than the step-wise procedure in the vast majority of cases. By contrast, in extreme cases of highly variable prices the simultaneous model may perform

slightly worse, if the bandwidths of prices used in the model do not sufficiently reflect that variability. In the other extreme case of perfectly known prices, the simultaneous model, again, performed substantially better than the simple step-wise approach, which clearly shows the need for a better representation of price uncertainties in such cases.

For this, in particular two strategies appear promising: Firstly, the use of more information on prices than just lower and upper bounds allowing for a more complex representation of price differences. The second option consists in increasing the level of detail of the root classification, in order to treat subgroups with prices of very different magnitudes separately (e.g., paper clips and pressure vessels of nuclear power plants, which both belong to fabricated metal products).

The extensive literature on *maximum entropy* models offers a wide variety of possible further extensions to the simultaneous estimation model. In this paper, external data used for the estimation of commodity flows and prices are treated as real numbers. However, they are themselves random variables, since any datum is the output of a series of surveys, projections and estimation procedures and, thus, always subject to uncertainty. This problem becomes particularly apparent in cases of information conflicts in external data. In such cases no feasible solution of the estimation problem exists. For such situations the concept of *generalized entropy* (see Golan et al., 1996, Robinson et al., 2001 and Rodrigues, 2014) or the treatment of information conflicts in the KRAS (Lenzen et al., 2009) algorithm may offer conceptual solutions.

The maximum entropy model developed in this chapter, offers the flexibility to be applied in any task, where flows measured in various units are to be estimated from partial information. Apart from the context of interregional commodity flows of this chapter, another important field of application constitutes combining monetary and physical I–O data. Several authors have shown that differences in the results of environmental impact assessments based on physical I–O tables versus environmentally extended monetary I–O can be attributed to issues that could be solved by the simultaneous estimation model developed here. These include the assumption of homogenous commodity prices across all uses (Weisz and Duchin, 2006), satisfying simultaneous financial, mass and/or energy balances (Merciai and Heijungs, 2014), as well as aggregating heterogeneous products consumed in different proportions by the various users (Majeau-Bettez et al., 2016).